

RNA-Seq Demo on Galaxy

Tom Doak

Le-Shin Wu

Carrie Ganote

National Center for Genome Analysis Support

July 16, 2014



INDIANA UNIVERSITY



INDIANA UNIVERSITY

Our RNA-Seq Demo Data



Cristobal Rojas, La miseria (1886)

We will be assembling the DNA Polymerase protein units from the H37Rv strain of *Mycobacterium tuberculosis*, the causative agent of TB, also known as the consumption.

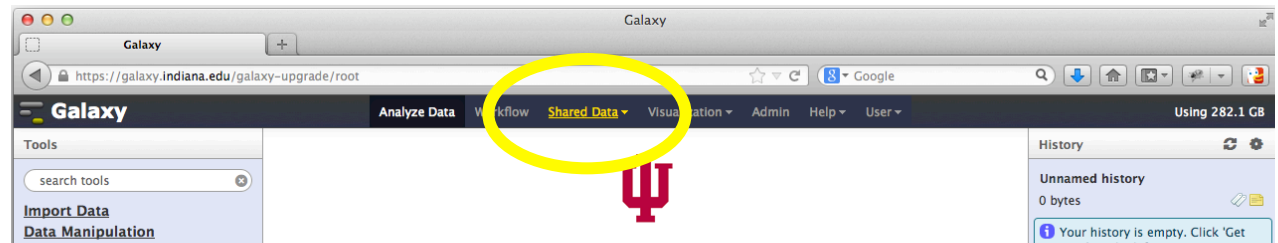
The raw reads originated from the Short Read Archive on NCBI. The accession number for the set is SRX212035.

This dataset consists of paired-end, ~75bp RNA-Seq reads.

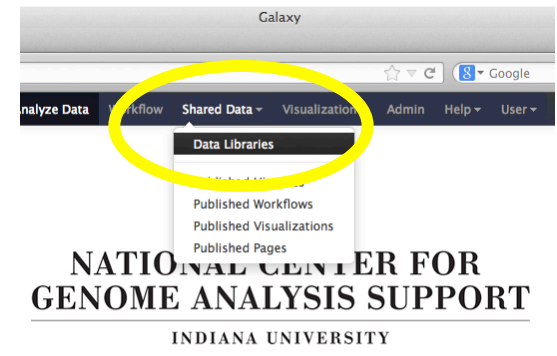


Let's get some sequence data

Galaxy allows users to publish their data to share with each other.



Let's start with "Shared Data" at the top.
Then select Data Libraries from the menu.





Let's get some sequence data

The screenshot shows the Galaxy web interface. At the top is a navigation bar with the Galaxy logo and links for 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', and 'Admin'. Below this is the 'Data Libraries' section, which includes a search bar with the placeholder text 'search dataset name, info, message, dbkey' and a magnifying glass icon. Below the search bar is a link for 'Advanced Search'. A table with two columns, 'Data library name' and 'Data library description', is displayed. The first row is 'User Import Library' with the description 'For moving large datasets into Galaxy'. The second row is 'Workshop Data' with the description 'Learning sets of RNA-Seq data'. The 'Workshop Data' link is circled in yellow.

<u>Data library name</u> ↓	<u>Data library description</u>
<u>User Import Library</u>	For moving large datasets into Galaxy
<u>Workshop Data</u>	Learning sets of RNA-Seq data

Choose Workshop Data.



Let's get some sequence data

Expand folder →

Check both boxes →

Name	Message	Data type
Galaxy Workshop September '13		
<input checked="" type="checkbox"/> TB_1.fq		fastqsanger
<input checked="" type="checkbox"/> TB_2.fq	Right reads	fastqsanger

For selected datasets:

TIP: You can download individual library datasets by selecting "Download this dataset" from the context menu (triangle icon).

TIP: Several compression options are available when downloading multiple library datasets simultaneously:

- gzip: Recommended for fast network connections
- bzip2: Recommended for slower network connections (smaller size but takes longer to compress)
- zip: Not recommended but is provided as an option for those who cannot open the above formats

Import the Data sets to current history.



Let's get some sequence data

Galaxy Analyze Data Workflow Shared Data Visualizations

Data Library "Workshop Data"

✓ 2 datasets imported into 1 history: Unnamed history

Name	Message
Galaxy Workshop September '13	
TB_1.fq	
TB_2.fq	Right reads

For selected datasets: Import to current history Go

i TIP: You can download individual library datasets by selecting "Download this dataset"

Data set is imported – Click on Analyze Data to return.



Step 1: Assess the Quality of Inputs

We will first get an idea of the quality of our input data sets.

The FastQC tool will produce graphical output that makes it easy to gauge the characteristics of the data – quality, patterns, biases, gc content etc.

The screenshot shows the Galaxy web interface. On the left, the 'Tools' panel is visible with a search bar and a list of tool categories. Under 'Quality Control', the tool 'FastQC:Read QC reports using FastQC' is highlighted with a yellow circle. The main panel shows the configuration for 'FastQC:Read QC (version 0.51)'. It includes a dropdown for 'Short read data from your current history' set to '2: TB_2.fq', a text input for 'Title for the output file' with the value 'FastQC', and a 'Contaminant list' dropdown set to 'Selection is Optional'. An 'Execute' button is at the bottom. Below the button, the 'Purpose' section explains that FastQC aims to provide a simple way to do quality control checks on raw sequence data, offering a quick overview, summary graphs, and an export option for an HTML report.

Choose either the left or right reads. Compare the results with your neighbor.

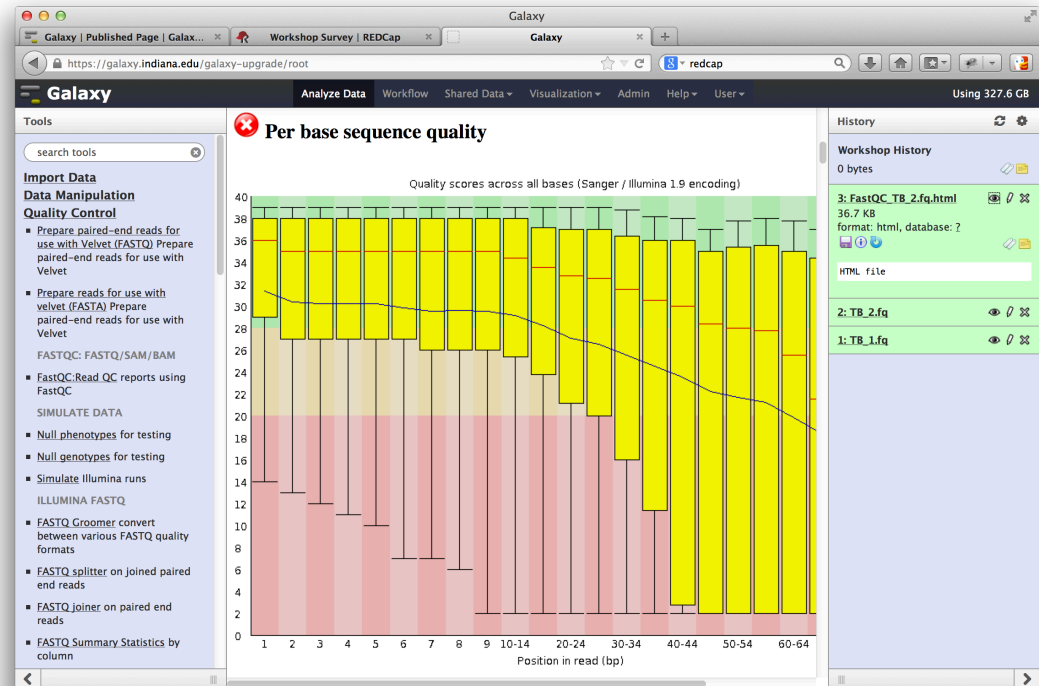


Step 1: Assess the Quality of Inputs

The input data usually declines in quality as the reads progress.

The quality score is assigned by the sequencing machine as it reads each base. It is a rough estimate of how ambiguous the signal is.

Sequence: **ATGCATG**
Quality Score: 39 38 23 19 3 3





Step 2: Trim Input Sequences

We've determined that the input data sets need some work before they are used in downstream processes. We'll use the FASTQ quality trimmer by sliding window to trim reads based on quality score.

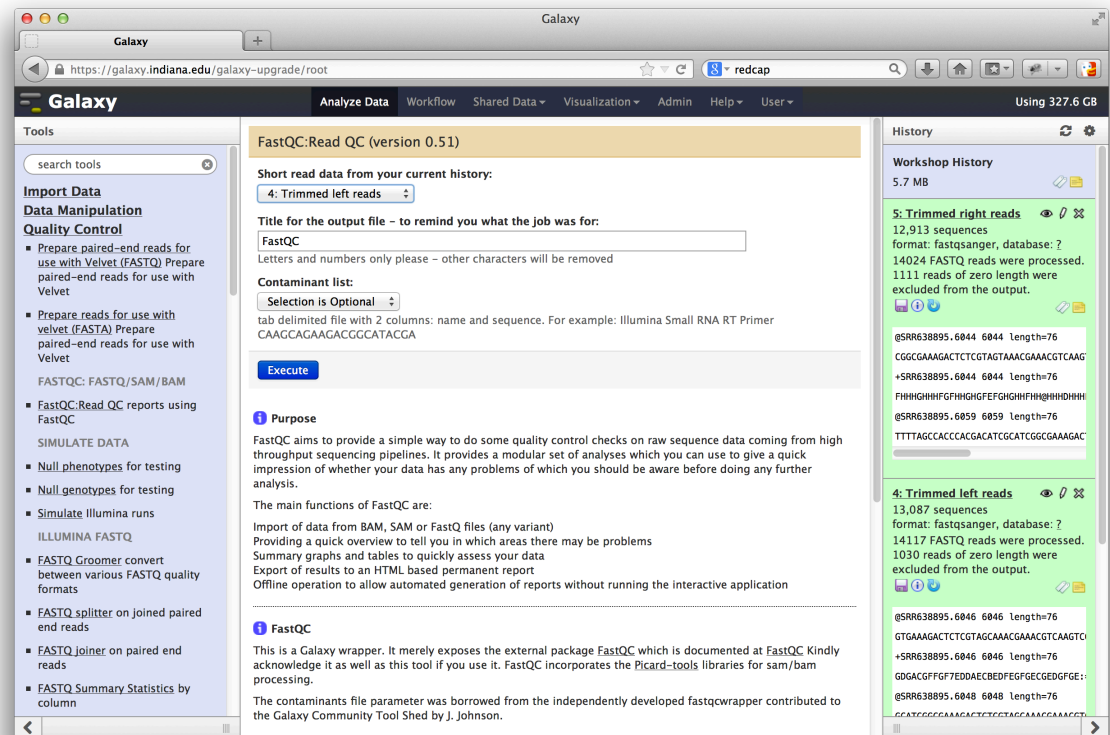
Run this tool for both input data sets.

The screenshot shows the Galaxy web interface. On the left, the 'Tools' panel lists various tools. The tool 'FASTQ Quality Trimmer by sliding window' is highlighted with a yellow circle. On the right, the tool's configuration page is shown. The title is 'FASTQ Quality Trimmer (version 1.0.0)'. The 'FASTQ File' is set to '1: TB_1.fq'. The 'Keep reads with zero length' checkbox is unchecked. The 'Trim ends' are set to '5' and 3''. The 'Window size' is set to 1. The 'Step Size' is set to 1. The 'Maximum number of bases to exclude from the window during aggregation' is set to 0. The 'Aggregate action for window' is set to 'min score'. The 'Trim until aggregate score is:' is set to '>='. The 'Quality Score' is set to 20.0. An 'Execute' button is at the bottom. Below the button, a description states: 'This tool allows you to trim the ends of reads based upon the aggregate value of quality score within a sliding window; a sliding window of size 1 is equivalent to 'simple' trimming of the ends of reads. The user specifies the aggregating action (min, max, sum, mean) to perform on the quality scores within the sliding window to be used with the user defined comparison operation and the quality score threshold.' The description is partially cut off at the bottom.



Step 3: Rinse, Repeat

Now that the files are trimmed, we will re-assess their quality. If necessary, keep trimming away until you are satisfied with the input files.



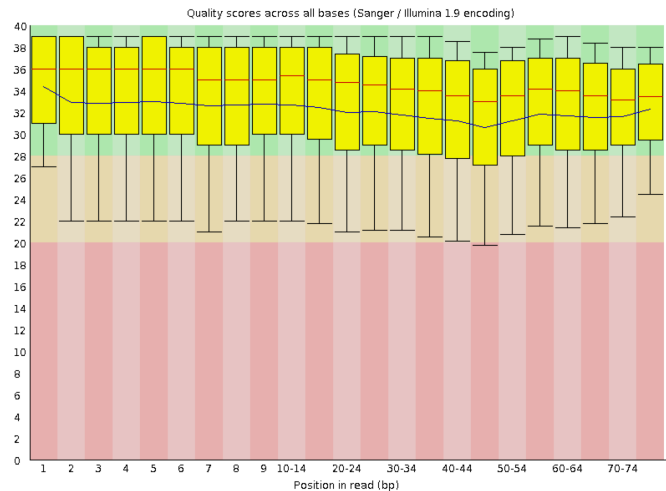
I renamed my trimmed files to help me keep them straight.



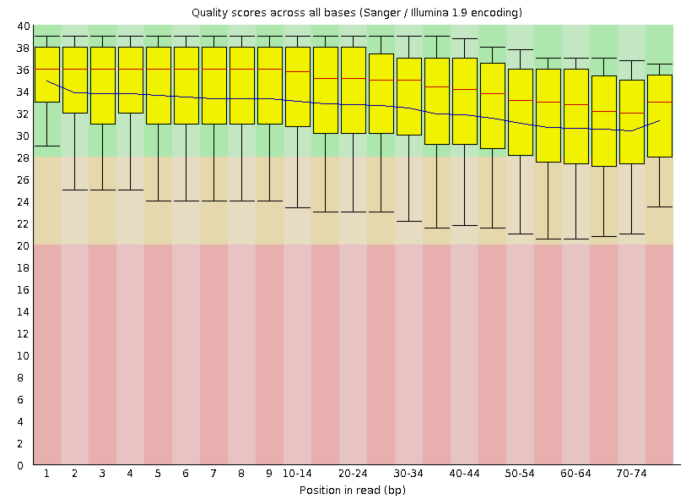
Step 3: Rinse, Repeat

Pictured are the left and right reads after trimming is complete.
These will do!

✓ Per base sequence quality



✓ Per base sequence quality





Step 4: Assembly

Next we will put the reads together to create a complete picture of the actively transcribed genes of the sample organism.

Trinity is a *de novo* assembler that has been optimized for use on Mason. We will use it to assemble our reads.

The screenshot displays the Galaxy web interface. On the left, a sidebar lists various tool categories: Tools, Import Data, Data Manipulation, Quality Control, De novo Assembly, Mapping and Alignments, Run Blast+, Run Blast+ on Open Science Grid, Annotation, Statistics, Variants, Clustering/Phylogeny, Visualization, and Workflows. The 'De novo Assembly' category is highlighted with a yellow oval, and the 'Trinity - Executes on Mason' tool is selected within it. The main panel shows the configuration for the Trinity tool (version 0.0.1). The configuration options include: 'Paired or Single-end data?' set to 'Paired'; 'Left/Forward strand reads' set to '4: Trimmed left reads'; 'Right/Reverse strand reads' set to '5: Trimmed right reads'; 'Strand-specific Library Type' set to 'None'; 'Paired Fragment Length' set to '300'; 'Is it strand specific data?' set to 'No'; 'Use Additional Params?' set to 'No'; and 'How long will your job need?' set to '1 hr'. An 'Execute' button is located at the bottom of the configuration panel.



It finished! We're done, right?

An assembler solves a computer problem of putting together a puzzle from tiny pieces. The output of the assembler is a guess – but we don't know how accurate it is. We could look at:

- Basic stats of the assembly – “Contigs”
 - Number of “Contigs” vs. Expected Number
 - N50 – a weighted average
 - Average Length
 - Max Length
- Check contigs against known genes with Blast

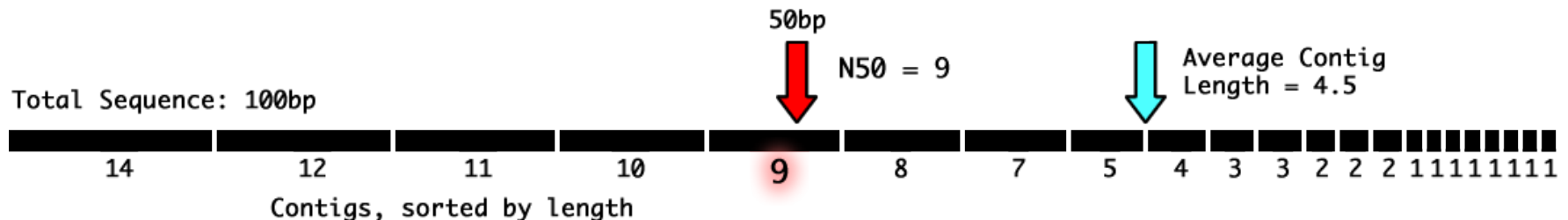
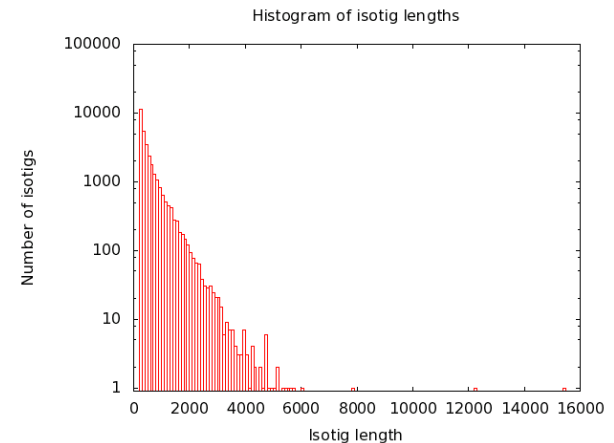


Step 5: Assessing Quality of Assembly

Important statistics for assembly quality:

Contig Length Distribution

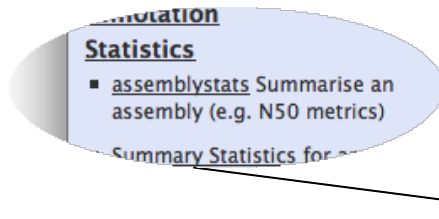
Assemblies will typically produce a number of complete contigs representing whole transcripts, and a large number of partial transcripts. This biases the average contig length toward the low end. The N50 is a measure weighted by total sequence length in the assembly.





Step 5: Assessing Quality of Assembly

Getting these stats in Galaxy:



Run assemblystats to get a summary and histograms of your contig length distribution.

The screenshot shows the Galaxy web interface with the 'assemblystats (version 1.0.1)' tool selected. The interface includes a top navigation bar with 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Admin', 'Help', and 'User'. The left sidebar lists various tools under 'Tools', with 'Statistics' highlighted. The main panel shows the tool's configuration: 'Type of read' is set to 'Isotig (if from transcriptomic assembly)', 'Output histogram with bin sizes=1' is checked, and the 'Source file in FASTA format' is '84: Trinity on data 20 and data 21: Assembled Transcripts'. An 'Execute' button is visible. The right sidebar shows the 'History' panel with two entries: '240: Sorted contigs' and '239: Assembly statistics'. The '239: Assembly statistics' entry is expanded, showing a table of statistics for isotig lengths.

Statistics for isotig lengths:
Min isotig length:
Max isotig length:
Mean isotig length:
Standard deviation of isotig leng
Median isotig length:



Step 6: Check Against Database

For this last step, we'll check to see how well our assembled transcripts compare to what we already know.

Use this step to give a rough annotation of genes, to make sure that your transcripts are from nuclear genes, or to gauge how complete your sequence is.

Galaxy Analyze Data Workflow Shared Data Visualization Admin Help

Tools

- SFF converter
- GTF-to-BEDGraph converter
- Wig/BedGraph-to-bigWig converter
- BED-to-bigBed converter
- FASTA**
- **Filter sequences by length**
- Compute sequence length
- Split FASTA alignment by species
- FASTA Width formatter
- RNA/DNA converter
- Collapse sequences

Filter sequences by length (version 1.1)

Fasta file:
9: Trinity - Executes on Mason on data 4 and data 5: Assemble

Minimal length:
0

Maximum length:
555
Setting to '0' will return all sequences longer than the 'Minimal length'

Execute

TIP. To return sequences longer than a certain length, set *Minimal length* to desired value and leave *Maximum length* set to '0'.

For sake of time, we'll just Blast one gene. Filter out to get the smallest.



Step 6: Check Against Database

We will use Blastx to search the NR database for our gene.

Use default search settings for this test set.

The screenshot shows the Galaxy web interface. On the left, the 'Tools' panel lists various tools. The tool 'NCBI BLAST+ blastx Search protein database with translated nucleotide query sequence(s)' is highlighted with a yellow circle. On the right, the configuration panel for 'NCBI BLAST+ blastx (version 0.0.17)' is shown. The 'Nucleotide query sequence(s)' is set to '16: Filter sequences by length on data 9'. The 'Subject database/sequences' is set to 'BLAST Database'. The 'Protein BLAST database' is set to 'NCBI NR (non redundant) 05 Jun 2012'. The 'Query genetic code' is set to '1. Standard'. The 'Set expectation value cutoff' is set to '0.001'. The 'Output format' is set to 'Pairwise HTML'. The 'Advanced Options' section is collapsed. An 'Execute' button is at the bottom.

Make sure to choose Pairwise HTML output for readability.



Step 6: Check Against Database

We see the expected genes as the top hits!

Analyze Data Workflow Shared Data Visualization Admin Help User Using 327.6 GB

Accession	Description
ref WP_009744290.1	DNA polymerase III subunit delta' [Actinomyces...
ref WP_005963184.1	DNA-directed DNA polymerase III subunit delta' [Actinomyces...
ref WP_010525702.1	DNA polymerase III subunit delta' [Nesterenkia...
ref WP_017201448.1	hypothetical protein [Microbacterium barkeri]
ref YP_830214.1	DNA polymerase III subunit delta' [Arthrobacter...
ref WP_019482356.1	hypothetical protein [Arthrobacter sp. TB 23]
ref WP_004806969.1	DNA polymerase III subunit delta' [Actinomyces...

```
>ref|WP_003907097.1| DNA polymerase III subunit delta, partial [Mycobacterium tuberculosis H37Rv]
gb|EFD75343.1| DNA polymerase III subunit delta [Mycobacterium tuberculosis H37Rv]
Length=354

Score = 246 bits (627), Expect = 3e-77
Identities = 181/181 (100%), Positives = 181/181 (100%), Gaps = 0/181
Frame = -3

Query 553 TDPQARQRRERALGLARDAATPSRAYAAAEELVAGAAEALALTAQRIEAEETEELRTA
Sbjct 174 TDPQARQRRERALGLARDAATPSRAYAAAEELVAGAAEALALTAQRIEAEETEELRTA

Query 373 aggtgkgtgaalrgatgAMKDLERRQKSRQTRASRDALDRALIDLATYFRDALLVAAH
Sbjct 234 AGGTGKGTGAALRGATGAMKDLERRQKSRQTRASRDALDRALIDLATYFRDALLVAAH

Query 193 GVRANHPDMADRVAAALAAHAPPERLLRCIEAVLACREALAVNVKPKFAVDAMVATIGC
Sbjct 294 GVRANHPDMADRVAAALAAHAPPERLLRCIEAVLACREALAVNVKPKFAVDAMVATIGC

Query 13 R 11
```

History

Workshop History

8.1 MB

17: blastx on db

484.1 KB
format: html, database: ?

HTML file

16: Filter sequences by length on data 9

1 sequences
format: fasta, database: ?

>comp7_c0_seq1 len=555 path=[1:0-555]
GCGGTCGCTACCGCAGTTCCTGGCCGATGGTGGCG
GTTTGACGTTGACCGCTAGCGCTTCCCTGCACGCC
GCCGCTCCGGCGGGGCGTGGGCGGCCAGCGCAGCA

We could limit the number of hits depending on output desired.



INDIANA UNIVERSITY

Step 7..?

RNA-Seq is a very versatile technology. You can use the data for:

- Gene discovery based on transcripts
- Genome evidence – introns, exons, junction
- Gene expression patterns
- SNP calling/other variants
- Protein divergence between samples

We have gotten to the assembly step, but there is a lot to learn about the data now that it is put together. A foundation in the use of Galaxy coupled with Indiana University resources will enable you to reach these goals.



INDIANA UNIVERSITY

Fin

Thanks for watching!
Questions and comments:
Email help@ncgas.org